Toward Automatic Fake News Classification

Souvick Ghosh
Rutgers University
souvick.ghosh@rutgers.edu

Chirag Shah
Rutgers University
chirags@rutgers.edu

Abstract

The interaction of technology with humans have many adverse effects. The rapid growth and outreach of the social media and the Web have led to the dissemination of questionable and untrusted content among a wider audience, which has negatively influenced their lives and judgment. Different election campaigns around the world highlighted how "fake news" - misinformation that looks genuine - can be targeted towards specific communities to manipulate and confuse them. Ever since, automatic fake news detection has gained widespread attention from the scientific community. As a result, many re-search studies have been conducted to tackle the detection and spreading of fake news. While the first step of such tasks would be to classify claims associated based on their credibility, the next steps would involve identifying hidden patterns in style, syntax, and content of such news claims. We provide a comprehensive overview of what has already been done in this domain and other similar fields, and then propose a generalized method based on Deep Neural Networks to identify if a given claim is fake or genuine. By using different features like the authenticity of the source, perceived cognitive authority, style, and content-based factors, and natural language features, it is possible to predict fake news accurately. We have used a modular approach by combining techniques from information retrieval, natural language processing, and deep learning. Our classifier comprises two main sub-modules. The first sub-module uses the claim to retrieve relevant articles from the know-ledge base which can then be used to verify the truth of the claim. It also uses word-level features for prediction. The second sub-module uses a deep neural network to learn the underlying style of fake content. Our experiments conducted on bench-mark datasets show that for the given classification task we can obtain up to 82.4% accuracy by using a combination of two models; the first model was up to 72% accurate while the second model was around 81% accurate. Our detection model has the potential to automatically detect and prevent the spread of fake news, thus, limiting the caustic influence of technology in the human lives.

1. Introduction

Many intellectuals have called the year 2016 as the beginning of a new era in modern politics, which has been named the "Post-truth" era. Subsequently, Oxford Dictionary[1] selected "Post-truth" as the Word of the Year ("English dictionary, translations and thesaurus," 2017) which is hardly surprising considering the trends observed globally, and the influence of fake news on electorates and public opinions.

According to a poll conducted by Pew Research Center in 2016[2] , Americans understand misinformation, and fake news constitute a significant societal problem, yet, most of them do not consider themselves responsible for the spread and dissemination of such information. Most of them strongly believe that they could detect false information when encountered and fewer agreed to have shared a fake news story with others.

With the increasing popularity and outreach of social media channels, such fake or misinformation could spread faster than ever before, reach a broader audience, and influence public opinion. Therefore, it has become increasingly important to address the concerns about fake news, using all possible approaches. Creating public awareness on how to judge a news item for veracity is one such approach, as is developing algorithmic methods to act as a first step in combating this ever-increasing problem.

As different disciplines of information science have been working towards mitigating this problem; it is essential to have a precise definition of the problem. What constitutes fake news? Cambridge dictionary

[1] https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016

[2] https://medium.com/trust-media-and-democracy/why-we-lie-to-ourselves-and-others-about-misinformation-770165692747

URI: https://hdl.handle.net/10125/59664
ISBN: 978-0-9981331-2-6

defines the above term as "false stories that appear to be news, spread on the Internet or using other media, usually created to influence political views or as a joke."[3] While we use this definition as a reference, fake news often involves the use of some other common terms like satire, propaganda, and rumor, which are also used for categorizing fake news. In our current research, we have tried to address the problem of fake news by developing a classifier which can automatically detect fake news accurately. We have trained our classifier on several benchmark datasets which comprise short sentences containing fake and real news obtained from several credible and fake sources (the details of the datasets have been provided in the Dataset Section).

Although many previous studies, many of them part of Fake News Challenge[4] , focused on automatic fake news detection, yet very few of them proposed universal models which could give acceptable results. A universal model should be able to correctly classify any type of claim by comparing it to factually supported information. This process should be independent of the source of the claim, which can be a blog post, tweet, mainstream media or oral speech.

The rest of the paper is organized as follows: Related Work section provides a thorough overview of the work which has already been done, and Dataset section describes the dataset. Sections Methodology and Experimental Results presents the experimental methodology and the results respectively. Conclusion and Future Work section concludes the paper and gives insight on the future scope of work.

2. Related Works

The massive popularity of social media has led to the availability of significant amount of user-generated, unregulated information which lacks in quality and are often unverifiable. Also, the content is generated in real-time in huge volumes (big data) and cannot be filtered or checked manually for veracity. This has resulted in the inundation of the Web with wrong or fake information - some of which are generated with malicious intent, and some for humor. Linguistically speaking, wrong information may be a result of inefficient reporting and may not be intended for misleading the audience or readers. However, the word fake involves planned actions for the purpose of presenting false information as true.

The rise of fake news in social media in recent years and the significant effects of it on the 2016 US elections, several studies have been conducted which relates to fake news, its influence and automatic detection. In this section, we try to mention and analyze the important researches which we find relatable to our work. We categorize them into two subsections: Traditional NLP approaches, and Deep Learning Approaches.

2.1. Traditional Natural Language Processing Approaches

Rubin, Chen, and Conroy (2015) [1] identified three types of fake news in their work. They categorized fake news into three distinct categories - serious fabrications, large-scale hoaxes, and humorous fake news. The ability of the social media like Facebook and Twitter to influence the opinions of audiences has led to increased use of fake information. This has made a significant impact on politics(voters) and e-commerce (online retailers).

Papadopoulou et al. (2017) [2] used a two-level text-based classifier to detect clickbaits. They used a wide variety of morphological, grammatical, stylistic, word-based features and sentiment analysis. Rubin et al. (2016) [3] used satirical cues to differentiate between fake and true news. Their approach depended on the absurdity of the text, punctuations, and grammatical features, and achieved a precision and recall of 90% and 87% respectively. Ahmed, Traore, and Saad (2017) [4] used Support Vector Machines with n-gram features in their work. They used tf-idf for feature extraction and linear SVM for the classification, achieving 92% accuracy on 50000 features.

Some researches adopted hybrid approaches by combining network analysis, sentiments, and behavioral information in addition to linguistic features. Conroy, Rubin, and Chen (2015) [5] were one of the first researchers to use network analysis in fake news detection while Mukherjee and colleagues (2013) [6] used words and the respective part-of-speech tags, together with bigrams to achieve a 68.1% accuracy on Yelp data. Bhelande et al. (2017) [7]used sentiment analysis using bag of positive and negative words for his Naive Bayesian classifier.

Researchers have also utilized discourse analysis with linguistics to identify instances of deception. Using language markers and rhetorical relations, Pisarevskaya (2017) [8] achieved an f-score of 0.65 using SVM and Random Forest classifiers.

2.2. Deep Learning Approaches

Shu and colleagues (2017) [9] provide a detailed overview of the recent approaches towards fake news detection and similar problems. While the problem of fake news detection is relatively new, there have been

[3]https://dictionary.cambridge.org/us/dictionary/english/fake-news
[4]http://www.fakenewschallenge.org/

several attempts to tackle it from an algorithmic (more specifically, machine learning) perspective. One such problem was proposed in the Fake News Challenge[5] (2017) where the participating teams were asked to detect the stance of the news claim.

One of the more famous problems of this kind was proposed in the Fake News Challenge (2017) where the participants tried to detect the stance of the claim. The organizers acknowledge that detecting the authenticity of a news story is a difficult and complex task, and hence, they reduced the original problem into a number of smaller problems, stance detection being one of them. Stance Detection focuses on evaluating a piece of news by understanding what other news organizations are saying about the same topic. Instead of evaluating a news claim as a standalone piece of information, it attempts to figure out the relative perspective of two pieces of text on a given topic or issue. Given a news article, the participants were required to classify the headline or claim as one of the following: agree, disagree, or irrelevant.

Many approaches have been investigated for solving this problem, which includes deep learning and traditional NLP techniques. Studying these approaches can be quite useful as they provide valuable insights into the problem at hand. Surprisingly, the top teams in the competition use simple but highly optimized methods to tackle the problem. For example, the second team (Riedel at al., 2017 [10]) and the third team ("Team Athene," 2017 [11]) used only simple multilayer Deep Neural Networks with highly optimized hyper-parameters and achieved accuracies of 85-88%. The former introduced a slightly more complicated approach by combining two classifiers, a deep learning model (made up of CNN layers and DL layers) and a gradient boosted tree classifier. Also, they use hand-made optimized features (Riedel et al. 2017 [10]).

Few other researches have adopted slightly more complicated approaches: modified versions of bidirectional LSTM/GRU architectures (Zeng, Zhou and Xu, 2017 [12]; Chopra, Jain and Sholar, 2017 [13]), ensemble of classifiers (Thorne et al., 2017 [14]), vanilla CNNs, independent encoders, conditional encoder (Rakholia and Bhargava, 2016 [15]), multipass conditional encoders, attentive readers with or without weighted cross entropy function (Miller and Oswalt, 2017 [16]) and bidirectional LSTMs. One team also treated the problem as a regression problem and introduced a new model called Siamese Regression model (Agarwal, Chin and Chen, 2017 [17])

Aymanns, Foerster, and Georg (2017) [18] treated the problem of fake news detection in social media as similar to finding distribution patterns in the social media graph. They used reinforcement learning and took into account if people supported or rejected the claim. Kumar (2017) [19] investigated the use of bots to spread fake news in social media and proposed a similar formulation of the problem. Avrahamov (2017) [20] constructed a knowledge-based graph by annotating each article with the information about its authors, topics, and main keywords. In their work, the problem is reduced to finding patterns in a hypergraph.

A similar problem is detecting rumors in tweets. Ma, Gao, and Wong (2017) [21] modeled the problem of classifying tweets (binary classification into either containing rumors or not) as a graph classification, by finding patterns in the distribution of tweet graph structure instead of checking the tweet text. In a separate work, Ma and colleagues (2016) [22] used RNN for classifying tweets as containing rumors or not. Jin et al. (2017) [23] addressed this problem by matching the tweets with verified articles which include rumors. Derczynski et al. (2017) [24] classified rumors in tweets into four categories using used ensemble methods, LSTMs and CNN. Chen et al. (2017) [25] have used a dataset of articles obtained from different sources of news, which could be fake or genuine. While the dataset was balanced, having an equal number of fake and reliable articles, they designed a three-layer hierarchical deep attentive reader with pooling to classify the test articles.

Researches focusing on identifying clickbaits provide useful insights on how to build automated systems to detect fake news. Many fake news posts are clickbaits, where the user is enticed to click on a given link. Biyani, Tsioutsiouliklis, and Blackmer (2016) [26] describes different types of clickbait posts and proposes a gradient boosted decision tree classifier to detect such clickbaits. Their model relies heavily on feature engineering, like the similarity in news headlines, informal nature of the posts and so on. Cao and Le (2017) [27] investigates different approaches using linear and logistic regressions, and random forests to detect clickbaits. Like Biyani and colleagues, Cao and Le use the tweet text and keywords for careful feature engineering.

Zhou (2017) [28] uses a self-attentive network with GRU cells for event-based Twitter/Weibo posts. Yang, Mukherjee and There are other researches which detect fake news using bidirectional LSTMs and external online sources (Karadzhov et al., 2017 [29]; Yang, 2017 [30]). Yang (2017) [30]identifies satirical news using Bidirectional RNN architectures with GRU cells and four levels of hierarchy augmented with attention

[5]Fake news challenge stage 1 (fnc-i): Stance detection, 2017. URL http://www.fakenewschallenge.org/.

mechanism. The dataset which they use contains articles labeled as satirical or real based only on the source of the news article.

Wang (2017) [31] introduced a new benchmark dataset, with six different categories, for fake news detection. This dataset contains metadata as well. Ruchansky, Seo, and Liu (2017) [32] adopt a multimodal approach by using the text, the associated images, and different social media features (e.g., the number of likes, shares, and tags for each post) for classification. For classification purposes, they used simple machine learning models like support vector machines (SVMs), random forest, and logistic regression.

3. Dataset

There is a lack of standard benchmark datasets for the problem of fake news detection; this is partly because the term fake news contains a wide variety of subcategories. Also, the scientific community has only been recently interested in tackling fake news, hence, the number of datasets developed solely for this purpose has been limited. The few datasets which are available publicly, differ from one another as they were designed for different types of tasks. We assessed over a dozen datasets which were used in related works. Out of these, only five of them were deemed relevant to our task of fake news detection.

We have further categorized the datasets into two types based on the length and structure of the sentences, the details of which are presented in the following subsections.

3.1. Type I Dataset

Type I datasets are for relatively short texts, as evidenced in case of tweets or news statements and headlines, which are typically 70 to 150 characters long. There were three different datasets which could be classified as Type I: LIAR dataset (Wang, 2017), Kaggle's Fake News Dataset (Risdal, 2016) and Fake News Challenge Dataset (Rubin, Chen, and Conroy, 2015).

3.1.1. LIAR This dataset was first introduced by Wang (2017) as a benchmark dataset for fake news detection problem. It contains around 12000 statements from various sources, each statement associated with a number which represents the truthfulness and credibility of the claim (the given statement) on a scale of 0 to 5 (0 being completely false and five being completely accurate). The statements and the labels were obtained from the Politifact Website which specializes in assessing the veracity of political statements. The assessment or labeling is done by expert journalists. Additionally, the dataset includes metadata information such as the speaker of the claim, position of the speaker, the home state if the speaker is a political representative, the history of his past statements and other similar information. The metadata information could be leveraged to detect an observable pattern in the way a person speaks. The dataset contains a large amount of news claims related to US politics and is considered hard to classify due to lack of sources or knowledge bases to verify with.

3.1.2. Kaggle's Fake News Dataset Kaggle.com[6] developed a dataset specific to fake news detection, which contains around 12500 instances (Risdal, 2016). Each instance is a claim which contains a header along with an article. The headlines of such article can be considered Type I while the text of the articles can be categorized as Type II. This dataset also contains some metadata such as crawl time and news id for each of the instances.

3.1.3. Fake News Challenge 2017 Dataset The Fake News Challenge Dataset (Rubin, Chen, and Conroy, 2015) contains around 13000 and 2587 full articles. Each instance contains a headline (which is mostly short), a reference to one of the articles, and the stance of the article towards the claim. The stance could be agree, disagree, discuss, or unrelated. Though the challenge approached the fake news through stance detection, which is unique and interesting, yet it requires a classification based on a pair of claim and article. In our work, we address the shortcoming by incorporating techniques from information retrieval and deep learning domains. In Figure 1, we present the frequency of the labels in the dataset. We also show the word cloud for this dataset in Figure 2, which gives some insight into the dominant topics in this dataset.

3.2. Type II Dataset

Type II datasets are made of longer texts, like what is observed in news articles, containing around 400 to 700 words. University of Washington Fake News Dataset was the only dataset which could be classified as Type II. The details of the datasets are presented in Table 1.

[6] https://www.kaggle.com/mrisdal/fake-news

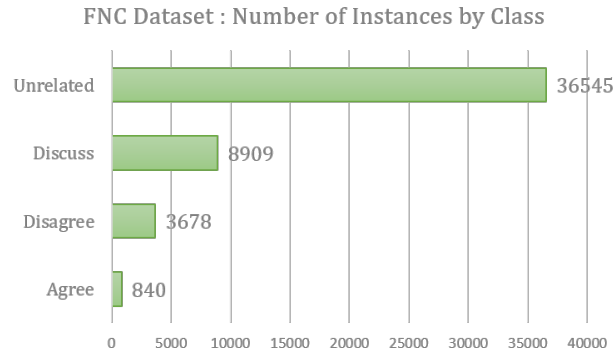

Figure 1: Label frequencies for Fake News Challenge Dataset

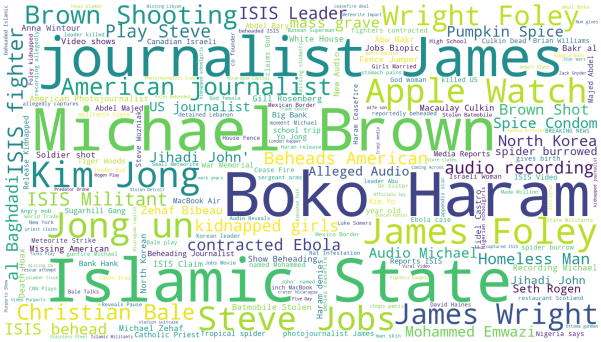

Figure 2: Word cloud for Fake News Dataset

3.2.1. University of Washington Fake News Dataset

This dataset boasts a total of 49000 instances, each comprising a paragraph of news article collected from credible and fake news sources (e.g., The Onion[7]). Each claim has one of four possible labels: hoax, propaganda, satire, or true news. The length of each sentence ranges between 500 to 600 words. Although the dataset was developed for a similar problem, we made slight modifications to make it more generalizable. For example, we removed all sentences which were labeled as satire as we theorize that satire is more of a linguistic phenomenon (intended for humor) than fake news (Rashkin et al., 2017 [33]).

Dataset	Avg. no. of instances	Avg. no. of words	Avg. no. of characters
Kaggle Fake News (text)	12999	637	NA
Kaggle Fake News (title)	12138	10.55	65
Fake News Challenge	49974	11	69
LIAR	12791	18	107
Univ. of Washington Fake News Data	60841	530	NA

Table 1: Comparison of Candidate Datasets

4. Methodology

To classify fake news, we have used a modular approach. The proposed model consists of several smaller submodules, each responsible for categorizing the instances based on a set of features. Finally, we combine the results through a voting process, which is based on a weighted average where the weights are also learned by the deep neural model. In this research, we have focused on two main submodules: the veracity detection submodule (based on information retrieval models and knowledge base) and the style based submodule. The main module can be extended by adding other submodules such as author metadata (background information, posting history, etc.) or cognitive authority of the source. In the following subsection, we discuss the details of the two main submodules which were implemented.

4.1. Veracity Detection Submodule

The first submodule is responsible for checking the veracity of each claim given that we have already constructed a knowledge base. To do so, two steps are taken: In the first step, the most relevant documents are retrieved from the knowledge base. In the second step, given those documents, the stance of the claim towards the documents is inferred. The overall flow of the process is depicted in Figure 4. This can be interpreted as checking the validity of a claim when a knowledge base of credible news sources is provided. The number of the retrieved documents is also controlled by a hand-picked hyper-parameter (denoted by k) of the model. It is evident that as we increase the hyperparameter k, the precision of the retrieved documents would suffer.

For retrieval, we used TF-IDF method as a baseline and more advanced algorithms for comparison and improved performance. The following three algorithms have been implemented and tested:

- BM25:
  BM25 algorithm (BM standing for Best Matching) is a ranking function scoring based

[7]https://www.theonion.com/

on probabilistic retrieval frameworks. It uses bag-of-words representation of documents to rank each document with respect to the different query words occurring in it. However, BM25 ignores the relative ordering of query terms as well as their proximity within the documents.

- Vector Space Model:
  The Vector Space Model is another retrieval algorithm which is implemented alongside the Boolean model of Information Retrieval in the Lucene framework. All the documents initially returned by the Boolean model are scored by the Vector Space Model and returned in ranked order. The ranking score is the cosine similarity between the query and the document vectors in a multidimensional word vector space. The advantages of this scoring method are partial matching and a continuous ranking scale.

- Language Model:
  This is another probabilistic model where conditional probability P(dlq) is calculated for the given query q and document d vectors. It assumes Dirichlet priors for the probability to smooth the function with a document normalization component.

After the k related articles are retrieved, in the second step of the algorithm, each article is classified into three labels 'Fake', 'Suspicious' or 'Legit.' For the classification, any deep learning architecture can be used. In our case, a simple Feed Forward Neural Net is used as shown in Figure 3. This specific architecture is inspired by one of three winning entries in Fake news challenge (Riedel et al., 2017 [10]). However, modifications are made to transform it to the reformulated problem.

The input features of the classifier are two one-hot bags-of-word vectors of size 5000, one corresponding to the news statement and the other to the article. Both vectors are fitted on the vocabulary of 5000 most frequently used words in the knowledge base. Additionally, it takes the cosine similarity between these two vectors as an additional input, hence, extending the final size of the input vector to 10001.

The hidden layer of the model has 100 Rectified Linear Units (ReLU), and the final layer is a SoftMax layer with three output classes as mentioned before.

4.2. Style Detection Submodule

The second submodule of our model is responsible for gaining valuable insights into how the writing style of fake news differs from real news. The syntax, semantics, and style of the written text can provide significant information about the intention of the authors. It has been widely observed that the language and tone of fake news presentation are more aggressive in general, and it involves a choice of words depicting strong emotions and biases (Rashkin et al., 2017 [33]). Our model uses a deep, bidirectional LSTM architecture. In past works, bidirectional LSTMs have proven efficient in storing, modeling and analyzing the information present in long sentences. The power of LSTMs come from their more complicated cell structures compared to standard RNNs. Also, using bidirectional neural networks instead of one-directional neural nets further improves the accuracy.

The equations for the LSTM model are as follows (Hochreiter and Schmidhuber, 1997 [34]):

$$\begin{aligned} i(t) &= \sigma(W^{(i)}x^{(t)} + U^{(i)}h^{(t-1)}) \\ f(t) &= \sigma(W^{(f)}x^{(t)} + U^{(f)}h^{(t-1)}) \\ o(t) &= \sigma(W^{(o)}x(^{t)}+U^{(o)}h^{(t-1)}) \\ \bar{c}^{(t)} &= tanh(W^{(c)}x^{(t)} + U^{(c)}h^{(t-1)}) \\ c^{(t)} &= f^{(t)} \circ \bar{c}^{(t-1)} + i^{(t)} \circ \bar{c}^{(t)} \\ h^{(t)} &= o^{(t)} \circ tanh(c^{(t)}) \end{aligned} \tag{1}$$

5. Experimental Results

To train our deep neural model, we used the Fake News Challenge (FNC) dataset to train the veracity-based (IR-DL) submodule, and the University of Washington Fake News Dataset (UW) to train the style-based module. One of the reasons for adopting this approach was the availability of knowledge base for the FNC dataset, and the richness of the UW dataset in terms of style (the other datasets focused on fact-based differences).

While training the veracity based module, the claims in the FNC dataset labeled as "unrelated" were discarded. For unrelated or irrelevant claims, there is no article which can help in verifying their authenticity (or the lack thereof), and therefore, it injects noise into the training.

As the average number of relevant documents in this dataset turned out to be 10, we chose the hyperparameter k to be 10. In Figures 5, 6, and 7, the recall and precision values for different values of k have been presented. By increasing the number k, which is the number of documents retrieved, the precision suffers, but recall improves. Table 2 represents the confusion matrix for the classification on FNC dataset.

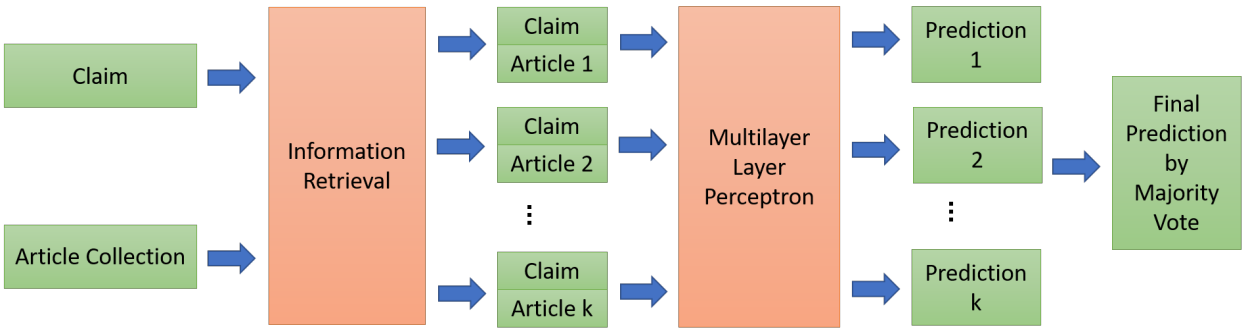

Figure 3: Overall Pipeline of the Veracity-based classifier

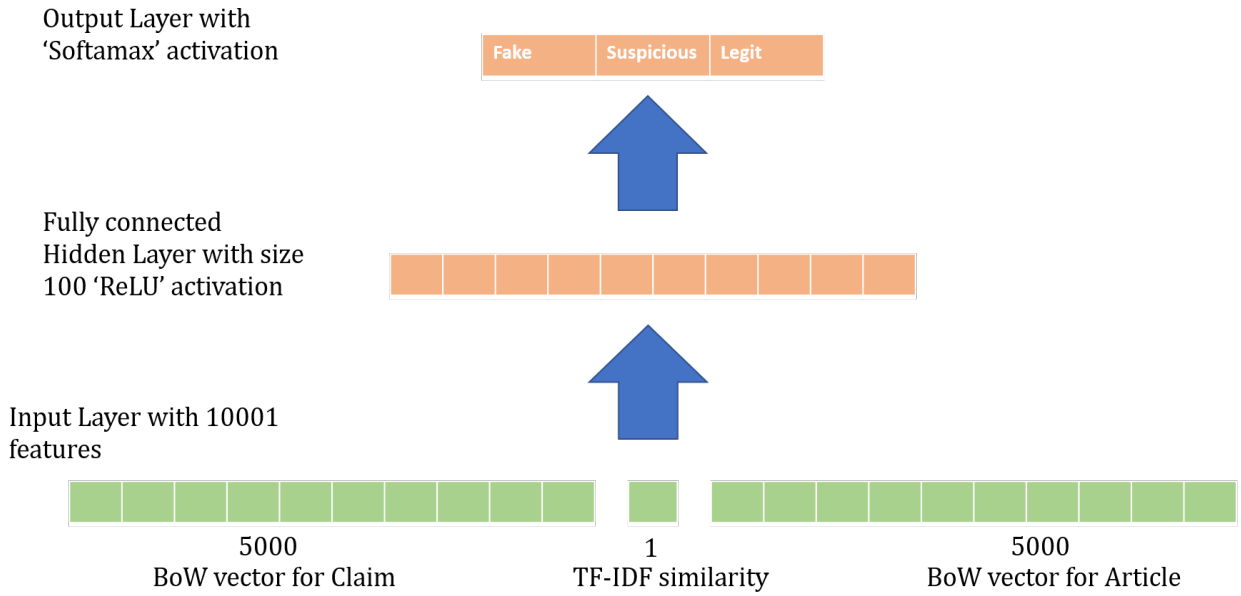

Figure 4: Architecture of the FFNN used

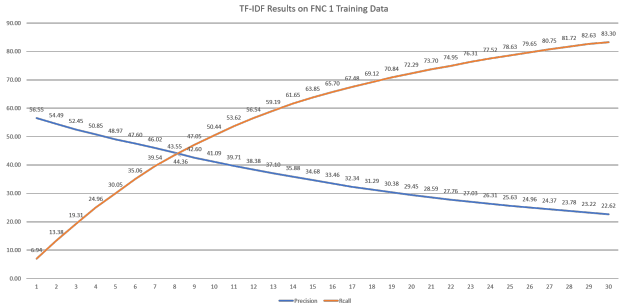

Figure 5: Precision and Recall for TF-IDF Method

The veracity-based submodule retrieved the most relevant documents relative to the claim and classified the claim into three possible mutually exclusive categories: fake, suspicious and real. The accuracy of prediction was 67.1% for ternary classification and 72.12% for binary classification. The style-based submodule, when evaluated separately on the UW test dataset, predicts with an accuracy of 81.83% (the best performing architecture).

Finally, by combining both the submodules using a weighted average, we were able to slightly increase the accuracy to 82.4%.

6. Discussions

Our work investigates on how to use techniques from the field of information retrieval and computer science to tackle the problem of fake news detection. While most of the previous works have focused on developing a machine learning classifier to address the problem, very few have considered using external knowledge base

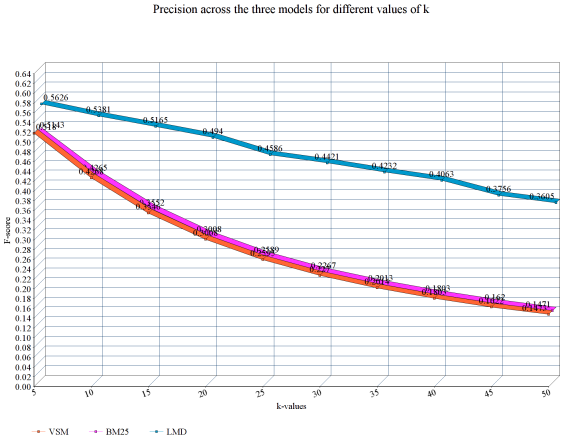

Figure 6: Precision for Advanced Algorithms

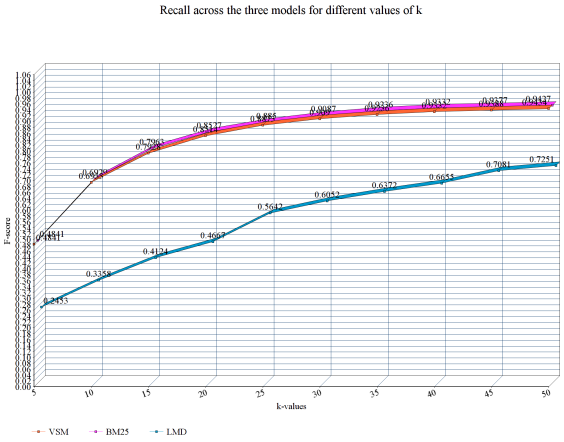

Figure 7: Recall for Advanced Algorithms

for improving the quality of the prediction. Our model, although limited by the datasets which are publicly available, tries to retrieve relevant articles which match the claim. To achieve this, we have used different information retrieval algorithms (like BM25, Vector Space Model, and Language Model). The algorithm could be extended to retrieve articles and documents from the Web which could lead to better understanding of the authenticity of the claim. We also develop a bidirectional LSTM model, which has shown admirable performance in tacking similar problems. By using different datasets for our work, we have also assessed the flexibility of the model for cross-domain analysis.

One of the significant difficulties faced while training and evaluating the veracity-based module was the class imbalance of the dataset. There were fewer instances of the class 'fake' which resulted in the biased training of the given classifier. To tackle this problem,

		Predicted		
		Suspicious	Fake	Legit
Actual	Suspicious	672	11	454
	Fake	138	12	173
	Legit	472	14	1889

Figure 8: Confusion Matrix for FNC Dataset

we used several approaches, such as merging datasets, oversampling or under-sampling the data. We also attempted to force the classifier to add extra penalty by modifying the cost function. This latter approach results in a higher precision and recall but lower accuracy (below 60%).

It should be noted that the performance of the IR-DL submodule (measured using accuracy) should not be compared to the Fake News Challenge (FNC) since FNC contains a considerable number of unrelated articles, which makes the task more manageable and the accuracy metric somewhat misleading. Owing to the unbalanced dataset, merely assigning each claim instance to the unrelated class would give an accuracy of 75%.

Our contribution is not limited to constructing an accurate model, but it advances the literature on fake news by evaluating how different retrieval techniques can be incorporated to deep neural architecture to create a more robust and flexible model. By modularizing the architecture, we allow for further enhancements and modules, such as the cognitive authority of source, mining of social media and public opinion and so on. However, our current model will need to be made more scalable to handle larger volumes of data.

7. Conclusion and Future Work

In this paper, we proposed a universal model to verify the authenticity of news claims. By using different features like the authenticity of the source, perceived cognitive authority, style, and content based factors, and natural language features, it is possible to accurately predict fake news. We have used a modular approach by combining techniques from information retrieval, natural language processing, and deep learning. Our classifier comprises two main submodules. The first submodule uses the claim to retrieve relevant articles from the knowledge base which can then be used to verify the truth of the claim. It also uses word-level features for prediction. The second submodule uses deep neural network to learn the underlying style of fake content. Our experiments

conducted on benchmark datasets show that for the given classification task we can obtain up to 82.4% accuracy by using a combination of two models; the first model was up to 72% accurate while the second model was around 81% accurate. Our detection model has the potential to automatically detect and prevent the spread of fake news, thus, limiting the caustic influence of technology in the human lives.

In the future, we would like to improve certain areas to improve the robustness of our model. One such improvement could be to modify the retrieval algorithm so that retrieval and learning are jointly performed, thus improving the accuracy. Also, we could use different architectures to evaluate if any of them outperform the architecture of our existing deep neural model. Few other submodules could also be constructed using author metadata (background information, posting history, etc.) or cognitive authority of the source. Another approach could be constructing a hypergraph of the authors and their articles and model a deep neural network on the graph.

One major problem that we faced was the lack of a generalized and standard dataset for the task of fake news detection. In the future, we would like to merge the existing datasets to create a universal benchmark dataset, with binary (fake or not fake) or ternary classification (fake, not fake, unsure) schemes, which could be used for fake news research. Lastly, we also intend to perform a qualitative evaluation of the different types of features that people perceive to be significant indicators of fake news. By constructing a human-centered theoretical model for fake news detection, we could advance the literature and lay the groundwork for future researches.

8. Acknowledgement

We would like to thank Dr. Gerard de Melo whose input has been helpful while designing the model, and Sepehr J. and Kshitij Shah, both of who contributed in initial conceptualization of the neural model.

References

[1] V. L. Rubin, Y. Chen, and N. J. Conroy, "Deception detection for news: three types of fakes," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1–4, 2015.

[2] O. Papadopoulou, M. Zampoglou, S. Papadopoulos, and I. Kompatsiaris, "A two-level classification approach for detecting clickbait posts using text-based features," *arXiv preprint arXiv:1710.08528*, 2017.

[3] V. L. Rubin, N. J. Conroy, Y. Chen, and S. Cornwell, "Fake news or truth? using satirical cues to detect potentially misleading news.," in *Proceedings of NAACL-HLT*, pp. 7–17, 2016.

[4] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pp. 127–138, Springer, 2017.

[5] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1–4, 2015.

[6] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What yelp fake review filter might be doing?," in *ICWSM*, 2013.

[7] M. Bhelande, A. Sanadhya, M. Purao, A. Waldia, and V. Yadav, "Identifying controversial news using sentiment analysis," *Imperial Journal of Interdisciplinary Research*, vol. 3, no. 2, 2017.

[8] D. Pisarevskaya, "Deception detection in news reports in the russian language: Lexics and discourse," in *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pp. 74–79, 2017.

[9] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[10] B. Riedel, I. Augenstein, G. P. Spithourakis, and S. Riedel, "A simple but tough-to-beat baseline for the fake news challenge stance detection task," *arXiv preprint arXiv:1707.03264*, 2017.

[11] A. Hanselowski, "Team athene on the fake news challenge - andreas hanselowski - medium," Aug 2017.

[12] S. X. Qi Zeng, Quan Zhou, "Neural stance detectors for fake news challenge," tech. rep., Stanford Uiversity, year = 2017.

[13] S. Chopra and S. Jain, "Towards automatic identification of fake news: Headline-article stance detection with lstm attention models," 2017.

[14] J. Thorne, M. Chen, G. Myrianthous, J. Pu, X. Wang, and A. Vlachos, "Fake news stance detection using stacked ensemble of classifiers," in *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pp. 80–83, 2017.

[15] S. B. Neel Rakholia, ""is it true?" - deep learning for stance detection in news," tech. rep., Stanford Uiversity, year = 2017.

[16] A. O. Kurt Miller, "Fake news headline classification using neural networks with attention," tech. rep., California State University, year = 2017.

[17] K. C. Akshay Agrawal, Delenn Chin, "Cosine siamese models for stance detection," tech. rep., Stanford University, year = 2017.

[18] C. Aymanns, J. Foerster, and C. Georg, "Fake news in social networks," *CoRR*, vol. abs/1708.06233, 2017.

[19] S. Kumar, "Investigating the use of bots to spread fake news in social media,"

[20] N. Avrahamov, "Machine learning, graphs, and the fake news epidemic (part 2) - dzone ai," Sep 2017.

[21] J. Ma, W. Gao, and K.-F. Wong, "Detect rumors in microblog posts using propagation structure via kernel learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 708–717, 2017.

[22] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks.," in *IJCAI*, pp. 3818–3824, 2016.

[23] Z. Jin, J. Cao, H. Guo, Y. Zhang, Y. Wang, and J. Luo, "Rumor detection on twitter pertaining to the 2016 us presidential election," *arXiv preprint arXiv:1701.06250*, 2017.

[24] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga, "Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours," *arXiv preprint arXiv:1704.05972*, 2017.

[25] T. Chen, L. Wu, X. Li, J. Zhang, H. Yin, and Y. Wang, "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," *arXiv preprint arXiv:1704.05973*, 2017.

[26] P. Biyani, K. Tsioutsiouliklis, and J. Blackmer, ""8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality.," 2016.

[27] X. Cao, T. Le, and J. Zhang, "Machine learning based detection of clickbait posts in social media," *CoRR*, vol. abs/1710.01977, 2017.

[28] Y. Zhou, "Clickbait detection in tweets using self-attentive network," *CoRR*, vol. abs/1710.05364, 2017.

[29] G. Karadzhov, P. Nakov, L. Màrquez, A. Barrón-Cedeño, and I. Koychev, "Fully automated fact checking using external sources," *CoRR*, vol. abs/1710.00341, 2017.

[30] F. Yang, A. Mukherjee, and E. Dragut, "Satirical news detection and analysis using attention mechanism and linguistic features," *arXiv preprint arXiv:1709.01189*, 2017.

[31] W. Y. Wang, ""liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv:1705.00648*, 2017.

[32] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 797–806, ACM, 2017.

[33] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2921–2927, 2017.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.